ORNL/CP-97084

To Fuse Or Not To Fuse: Fuser Versus Best Classifier †

Nageswara S. V. Rao
Mailstop 6355, Bldg. 6010
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-6355
email: raons@ornl.gov

To be presented at *SPIE Conference on Sensor Fusion: Architecture and Applications*, April 13-17, 1998, Orlando, Florida.

---

†Research sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation, the Seed Money Program of Oak Ridge National Laboratory, and the Office of Naval Research under order N00014-96-F-0415.

DISCLAIMER

CONF-980412--

To fuse or not to fuse: Fuser versus best classifier

Nageswara S. V. Rao

Oak Ridge National Laboratory, Oak Ridge, TN 37831-6355,USA

ABSTRACT

A sample from a class defined on a finite-dimensional Euclidean space and distributed according to an unknown distribution is given. We are given a set of classifiers each of which chooses a hypothesis with least misclassification error from a family of hypotheses. We address the question of choosing the classifier with the best performance guarantee versus combining the classifiers using a fuser. We first describe a fusion method based on isolation property such that the performance guarantee of the fused system is at least as good as the best of the classifiers. For a more restricted case of deterministic classes, we present a method based on error set estimation such that the performance guarantee of fusing all classifiers is at least as good as that of fusing any subset of classifiers.

Keywords: Classification, finite sample analysis, distributed detection, fusion of classifiers

1. INTRODUCTION

Over the past decades several methods, such as nearest neighbor rules, neural networks, tree methods, and kernel rules, have been developed for designing classifiers. Often, the classifiers are quite varied and their performances are characterized by various smoothness and/or combinatorial parameters.[1] The designer is thus faced with a wide variety of choices which are not easily comparable. It is generally known that a good fuser outperforms the best classifier, and at the same time, a bad fuser choice can result in a performance worse than the worst classifier. Thus it is very important to employ fusion methods that provide concrete performance guarantees – in particular, for the fuser to be meaningful it must perform at least as well as the best classifier. If the underlying joint distributions are known, the classifiers can be combined optimally by using available distributed detection methods.[2] In the special case of statistically independent classifiers, one can employ linear combinations to combine outputs of classifiers.[3] In practical classifier systems, however, independence is seldom satisfied, and the underlying distributions are very hard to estimate since sample is often the only information available. Although the theory of sample-based classifier design has been well developed,[1] an analogous theory for fusion of classifiers is developed only to a limited extent. In this paper, we describe two fusion methods that are applicable to sample-based fusion of multiple classifiers. We restrict our attention to the classifiers for which distribution independent performance guarantees can be provided. This formulation is based on Vapnik and Chervonenkis theory,[4,5] which has been extensively studied recently in the *probably approximately correct* (PAC) learning paradigm.[6,7]

A classical pattern recognition problem is stated as follows: we are given an independently and identically distributed (iid) sample (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), according to an unknown distribution $P_{X,Y}$, where $X_i \in \Re^d$ and $Y_i \in \{0, 1\}$. The problem is to design a classifier $\phi : \Re^d \mapsto \{0, 1\}$ based on the sample that ensures a small value for the *probability of misclassification* given by

$$L(\phi) = \int_X I_{\{\phi(X) \neq Y\}} dP_{X,Y},$$

where $I_D(x)$ is the *indicator function* of the set $D \subseteq \Re^d$ such that $I_C(x) = 1$ if $x \in C$ and $I_C(x) = 0$ otherwise. We often suppress the operand x when it is clear from the context.

In the formulation based on Vapnik-Chervonenkis theory,[5,1] ϕ is chosen from a class $\mathcal{H}$. Since $P_{X,Y}$ is unknown, exact minimization of $L(.)$ is not possible. Instead, we consider the *empirical error of misclassification* given by

$$\hat{L}(\phi) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\phi(X_i) \neq Y_i\}}.$$

Other author information: E-mail: raons@ornl.gov

Let $\hat{\phi}$ minimize $\hat{L}(.)$ over $\mathcal{H}$. If $\mathcal{H}$ has finite Vapnik-Chervonenkis dimension $V_{\mathcal{H}}$, it is well-known[1] that one can guarantee

$$P^n_{X,Y}\left[L(\hat{\phi}) - \min_{\phi \in H} L(\phi) > \epsilon\right] \leq \delta$$

if n is chosen to be sufficiently large, irrespective of the distribution $P_{X,Y}$. This condition asserts that the misclassification error committed by $\hat{\phi}$ is within ϵ of the best possible error, namely $\min_{\phi \in H} L(\phi)$, with a probability of at least $1 - \delta$.

We are given N such classifiers corresponding to the classes $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_N$ such that

$$P^n_{X,Y}\left[L(\hat{\phi}_i) - \min_{\phi \in H_i} L(\phi) > \epsilon\right] \leq \delta_i$$

where $\hat{\phi}_i$ minimizes $\hat{L}(.)$ over $\mathcal{H}_i$. If the classifiers are statistically independent, it is well-known that the higher the number of classifiers the better is the performance of the fused system.[8] Such result is not true if independence is not satisfied. Our objective is to "fuse" the classifiers so that the fused system performs at least as well as the best individual classifier based on the sample only. If the joint error distributions of the classifiers are known, then the fusion problem can be solved using the existing maximum likelihood estimation methods.[2] The main challenge of the present formulation is due to the lack of knowledge of error distributions. Problems of this kind are of relatively recent interest with most works dealing with computing a close-to-optimal fusion rule within a class[9] or sample-based implementation of fusion rules derived for known distributions case.[10] In particular, this is a special case of the generic sensor fusion problems studied recently.[11-13] Very few results exist for the present problem that guarantee that the fused system is at least as good as the best classifier or best combination of classifiers (with some exceptions[14,15]).

In this paper, we describe two methods that enable us to judge the performance of the fused system. The first method is based on the isolation property[16] that enables us to compare the fused system with the best individual classifier. This method is simple to apply and requires easily satisfiable criteria. The second method is based on intersections of error sets of the classifiers, and enables us to decide the relative performance of the fused system in comparison with any subset of classifiers. This method requires more stringent conditions.

2. SINGLE CLASSIFIER

We now summarize the known results for a single classifier.[1] The lowest possible error achievable by any deterministic classifier is given by the *Bayes error* $L(\phi^*)$, where $\phi^* : \Re^d \mapsto \{0, 1\}$ is defined as

$$\phi^*(x) = \begin{cases} 1 & \text{if } P_{X,Y}[Y = 1 | X = x] \geq P_{X,Y}[Y = 0 | X = x] \\ 0 & \text{otherwise} \end{cases}$$

Since the distribution is not known, ϕ^* cannot be computed. Furthermore, based on a finite sample, only an approximation to $L(\phi^*)$ can be achieved in general. In particular, the performance of $\hat{\phi}$ that minimizes $\hat{L}(.)$ can be characterized by the properties of $\mathcal{H}$.

Let $\mathcal{A}$ be a collection of measurable sets of R^d. For $(z_1, z_2, \ldots, z_n) \in \{\Re^d\}^n$, let $\mathcal{N}_{\mathcal{A}}(z_1, z_2, \ldots, z_n)$ denote the number of different sets in

$$\{\{z_1, z_2, \ldots, z_n\} \cap A : A \in \mathcal{A}\}.$$

The nth *shatter coefficient* of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(z_1, z_2, \ldots, z_n) \in \{\Re^d\}^n} \mathcal{N}_{\mathcal{A}}(z_1, z_2, \ldots, z_n).$$

Then, the *Vapnik-Chervonenkis* (VC) dimension of $\mathcal{A}$, denoted by $V_{\mathcal{A}}$, is the largest integer $k \geq 1$ such that $s(\mathcal{A}, k) = 2^k$. The following important identity[5,7] relates the shatter coefficient to VC dimension:

$$s(\mathcal{A}, n) = \begin{cases} 2^n & \text{if } n \leq V_{\mathcal{A}} \\ 2\frac{n^{V_{\mathcal{A}}}}{V_{\mathcal{A}}!} & \text{if } n > V_{\mathcal{A}} \end{cases}$$

Then we have the following result

$$P^n_{X,Y}\left[\sup_{\phi\in\mathcal{A}}|\hat{L}(\phi)-L(\phi)|>\epsilon\right]\leq 8s(\mathcal{A},n)e^{-n\epsilon^2/32}.$$

which in turn implies that

$$P^n_{X,Y}\left[L(\hat{\phi})-\min_{\phi\in H}L(\phi)|>\epsilon\right]\leq 8s(\mathcal{H},n)e^{-n\epsilon^2/128}.$$

Thus, given a sample of size

$$n=\frac{128}{\epsilon^2}\left(\ln s(\mathcal{H},n)+\ln(8/\delta)\right)$$

we have

$$P^n_{X,Y}[L(\hat{\phi})-\min_{\phi\in\mathcal{H}}L(\phi)>\epsilon]<\delta,$$

irrespective of the distribution $P_{X,Y}$.

3. ISOLATION PROPERTY

We consider a family of fuser functions $\mathcal{F}:\{f:\{0,1\}^N\mapsto\{0,1\}\}$ such that the fused output is given by $f[\hat{\phi}_1(X),\hat{\phi}_2(X),\ldots,\hat{\phi}_N(X)]$, denoted by $f(Z)$, where $Z=(\hat{\phi}_1(X),\hat{\phi}_2(X),\ldots,\hat{\phi}_N(X))$. The error probability of the fused system is given by

$$L_F(f)=\int I_{\{f(Z)\neq Y\}}dP_{X,Y}.$$

Note that Z is a deterministic function of X given the sample. For computational convenience, we utilize the following alternative formula

$$L_F(f)=\int[f(Z)-Y]^2dP_{X,Y}.$$

Note that $|\mathcal{F}|\leq 2^{2^N}$ since $\mathcal{F}$ consists of at most all Boolean functions on N variables. Consider the function class

$$\mathcal{G}=\{f(\phi_1(X),\phi_2(X),\ldots,\phi_N(X)):\phi_1\in\mathcal{H}_1,\phi_2\in\mathcal{H}_2,\ldots,\phi_N\in\mathcal{H}_N\}.$$

Here $f(\phi_1(.),\phi_2(.),\ldots,\phi_N(.))$ specifies a subset of $\Re^d$, and hence $\mathcal{G}$ specifies a family of sets of $\Re^d$.

The fuser is obtained in two steps: (a) a training set (Z_1,Y_1), (Z_2,Y_2), ..., (Z_n,Y_n), where $Z_i=(\hat{\phi}_1(X_i),\hat{\phi}_2(X_i),\ldots,\hat{\phi}_N(X_i))$, is derived from the classifiers and the original sample, and (b) the fuser is derived by minimizing empirical error over $\mathcal{F}$. Let f^* minimize $L_F(.)$ over $\mathcal{F}$. Note that f^* cannot be exactly computed since $P_{X,Y}$ is unknown. Instead, we minimize the empirical error given by

$$\hat{L}_F(f)=\frac{1}{n}\sum_{i=1}^{n}[f(Z_i)-Y_i]^2.$$

Let $\hat{f}$ minimize $\hat{L}_F(.)$ over $\mathcal{F}$.

If one of the classifier is to be chosen, the lowest achievable error is given by $\min_{i=1}^{N}L(\phi_i^*)$. Since the classifiers can be correlated in an arbitrary manner, the empirically best classifier $\hat{\phi}_{\min}=\arg\min_i\hat{L}(\hat{\phi}_i)$ yields the following guarantee

$$P^n_{X,Y}\left[L(\hat{\phi}_{\min})-\min_{i=1}^{N}L(\phi_i^*)>\epsilon\right]<\delta_1+\delta_2+\ldots+\delta_N.$$

The fuser, thus, provides a *better guarantee* if $\delta_F<\delta_1+\delta_2+\ldots+\delta_N$ where

$$P^n_{X,Y}\left[L_F(\hat{f})-\min_{i=1}^{N}L(\phi_i^*)>\epsilon\right]<\delta_F.$$

Definition 1. The fuser class $\mathcal{F}$ satisfies the *isolation property*[16,17] if it contains the following N functions: for all $i = 1, 2, \ldots, N$ we have $f_i(z_1, z_2, \ldots, z_N) = z_i$.

This property is trivially satisfied if $\mathcal{F}$ consists of all Boolean functions of N variables. Although it is sufficient to include N functions in $\mathcal{F}$ to satisfy this property, in general a richer class performs better in practice.

Theorem 1. If the fuser class $\mathcal{F}$ satisfies the isolation property, then fuser $\hat{f}$ provides better guarantee than the best classifier under the condition

$$|\mathcal{F}| \leq \frac{1}{2} \sum_{i=1}^{N} \delta_i e^{\epsilon^2 n/2}.$$

Proof: We first have

$$\begin{aligned} L_F(f^*) &= \min_{f \in \mathcal{F}} \int [f(Z) - Y]^2 dP_{X,Y} \\ &\leq \min_{i=1}^{N} \int [f_i(Z) - Y]^2 dP_{X,Y} \\ &\leq \min_{i=1}^{N} \int [\hat{\phi}_i(X) - Y]^2 dP_{X,Y} \\ &= \min_{i=1}^{N} \int I_{\{\hat{\phi}_i(X) \neq Y\}} dP_{X,Y} = \min_{i=1}^{N} L(\hat{\phi}_i) \end{aligned}$$

where the third step is a direct consequence of the isolation property. Consequently the event $\{L_F(\hat{f}) - \min_{i=1}^{N} L(\hat{\phi}_i) > \epsilon\}$ implies the event $\{L_F(\hat{f}) - L_F(f^*) > \epsilon\}$. Thus we have

$$\begin{aligned} & P_{X,Y}^n \left[L_F(\hat{f}) - \min_{i=1}^{N} L(\hat{\phi}_i) > \epsilon \right] \\ & \leq P_{X,Y}^n \left[L_F(\hat{f}) - L_F(f^*) > \epsilon \right] \\ & \leq 2|\mathcal{F}| e^{-\epsilon^2 n/2} \end{aligned}$$

where the last step is due to the finiteness of $|\mathcal{F}|$.[5] □

A minimal realization of this theorem can be based on $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ as per the isolation property defined above. We wish to emphasize that this fusion method can be easily applied without identifying the best classifier, while still ensuring its performance in the fused system. The condition of Theorem 1 can be expressed in terms of the VC dimensions as follows

$$|\mathcal{F}| \leq 4 \sum_{i=1}^{N} \frac{(n)^{V_{\mathcal{H}_i}}}{V_{\mathcal{H}_i}!} e^{-\epsilon^2 63n/128}.$$

by noting that $\delta_i = \frac{8(n)^{V_{\mathcal{H}_i}}}{V_{\mathcal{H}_i}!} e^{-\epsilon^2 n/128}$ for $n > \max(V_{\mathcal{H}_1}, V_{\mathcal{H}_2}, \ldots, V_{\mathcal{H}_N})$.[4]

4. ERROR SET INTERSECTIONS

We consider deterministic class in this section such that $Y = I_C(X)$ for some $C \in \Re^d$. In this case $\phi^* = C$ and $L(\phi^*) = L(C) = 0$. Furthermore, we have

$$L(\phi) = \int I_{\{\phi(X) \neq I_C(X)\}} dP_X = E_X[I_{\{\phi(X) \neq I_C(X)\}}],$$

where $E_X[.]$ denotes the expectation with respect to X.

The outline of this fusion method is as follows: (a) first estimate the sets over which the classifiers make errors, and (b) compute the fuser as the complement of the intersection of these sets. Thus the fuser makes an error only if all the classifiers make an error. The effectiveness of the method depends on the efficiency of the error set estimation method.

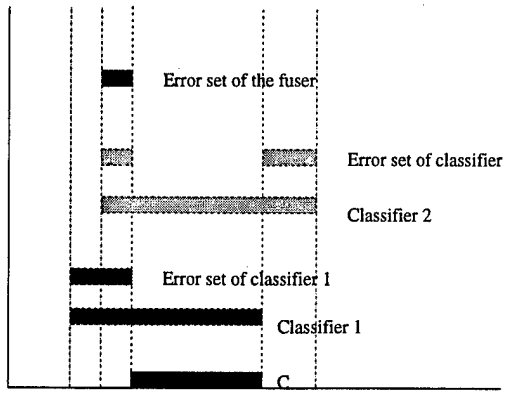


Figure 1. Illustration of intersection of error sets.

We describe our method in two steps. For the sake of explanation, we first assume that the ϕ_i^*'s and C are known. We then replace ϕ_i^* by $\hat{\phi}_i$ and estimate $\phi_i^* \oplus C$ by $\hat{E}_i$ chosen from a suitable family $\mathcal{E}_i$.

By denoting the set $\{x : \phi(x) = 1\}$ by ϕ itself (with an abuse of notation), the error set of ϕ is $\phi \oplus C$, where $\phi \oplus C$ is the symmetric difference given by $(\bar{\phi} \cap C) \cup (\phi \cap \bar{C})$. Under this notation, we have $L(\phi) = E_X[\phi \ominus C]$. The *ideal fuser* is given by

$$\tilde{f} = \overline{\bigcap_{i=1}^{N} (\phi_i^* \oplus C)} = \bigcup_{i=1}^{N} \overline{(\phi_i^* \oplus C)}$$

which implies that it makes an error if and only if all classifiers make an error. Thus we have $L(\tilde{f}) = E_X\left[\bigcap_{i=1}^{N} (\phi_i^* \oplus C)\right]$. Let B be a subset of $\{1, 2, \ldots, N\}$, and $\tilde{f}_B = \overline{\bigcap_{j \in B} (\phi_j^* \oplus C)}$, which is a fuser based on a subset of the classifiers. Since $\bigcap_{i=1}^{N} (\phi_i^* \oplus C) \subseteq \bigcap_{j \in B} (\phi_j^* \oplus C)$, we have

$$L(\tilde{f}) \leq E_X\left[\bigcap_{i=1}^{N} (\phi_i^* \oplus C)\right] \leq \min_{B \subseteq 2^{\{1,2,\ldots,N\}}} E_X\left[\bigcap_{j \in B} (\phi_j^* \oplus C)\right] = L(\tilde{f}_B).$$

In particular, we have

$$L(\tilde{f}) = E_X\left[\bigcap_{i=1}^{N} (\phi_i^* \oplus C)\right] \leq \min_{i=1}^{N} E_X[\phi_i^* \oplus C] = \min_{i=1}^{N} L(\phi_i^*).$$

Thus $\tilde{f}$ has a very important property: its performance is at least as good any subset of classifiers, i.e. one does not do better by considering a classifier subset; in particular, the fuser performs as well as the best classifier.

Example 1. To illustrate the main idea, let C correspond to a interval on real line as shown in Fig. 1. Let $\mathcal{H}_i$ consist of intervals such that the ϕ_i^* corresponds to an interval and $\phi_i^* \oplus C$ corresponds to union of at most two intervals as illustrated in Fig. 1. The intersection of error sets consists of unions of intervals whose total length is no larger than that of intersection of error sets of any subset of classifiers. In the figure this set consists of a single interval, and note that typically this interval is smaller in length than that of any classifier. □

Since ϕ_i^*'s and C are unknown, the ideal fuser $\tilde{f}$ cannot be computed. In place of ϕ_i^* we have its estimate given by the classifier $\hat{\phi}_i$. We then estimate for each classifier the *error set* given by $\hat{\phi}_i \oplus C$ by employing the class $\mathcal{E}_i$. Let $\hat{E}_i \in \mathcal{E}_i$ be an *empirically consistent* estimator of $\hat{\phi}_i \oplus C$ in that $\hat{L}(E_i) \leq \hat{L}(\hat{\phi}_i \oplus C)$. Then the fuser based on the sample is computed as $\overline{\bigcap_{i=1}^{N} \hat{E}_i}$.

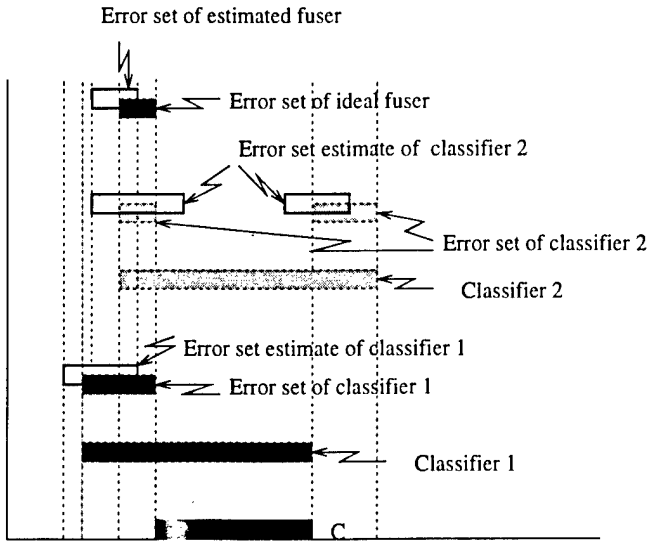


Figure 2. Illustration of fuser computation.

Example 2. Consider that C and P_X are unknown in the case of Example 1. In this case, $\mathcal{E}_i$ consists of unions of two intervals, and in general we can compute only an approximation $\hat{E}_i$ to the error set $\overline{\phi_i^* \oplus C}$ as shown in Fig. 2. The fuser is computed as the complement of intersection of of the error sets, given by $\overline{\bigcap_{i=1}^{N} \hat{E}_i}$, which in general only provides an approximation to the target $\overline{\bigcap_{i=1}^{N} (\phi_i^* \oplus C)}$. □

The performance of this fusion method depends on that of $\hat{E}_i$'s as characterized in the following theorem.

Theorem 2. Consider consistent error set estimators such that $\hat{L}(\hat{E}_i) \leq \hat{L}(\hat{\phi}_i \oplus C)$, for $i = 1, 2, \ldots, N$, where C is the target class, and $\hat{L}(\hat{\phi}_i) = \min_{\phi \in \mathcal{H}_i}{}^{N} \hat{L}(\phi)$. Then, we have

$$P_X^n \left[\left| L\left(\overline{\bigcap_{i=1}^{N} \hat{E}_i} \right) - L(\tilde{f}) \right| > \epsilon \right] \leq 8 \left[s\left(\bigcap_{i=1}^{N} \mathcal{E}_i, n \right) + s\left(\bigcap_{i=1}^{N} \mathcal{H}_i, n \right) \right] e^{-\epsilon^2 n/512}$$

where $\tilde{f} = \overline{\bigcap_{i=1}^{N} (\phi_i^* \oplus C)}$ is the ideal fuser, and ϕ_i^* is the best individual classifier, i.e. $L(\phi_i^*) = \min_{\phi \in \mathcal{H}_i} L(\phi)$.

Proof: By noting that $L\left(\bigcap_{i=1}^{N} \hat{E}_i \right) + L\left(\overline{\bigcap_{i=1}^{N} \hat{E}_i} \right) = 1$ and $L\left(\overline{\bigcap_{i=1}^{N} (\phi_i^* \oplus C)} \right) + L\left(\bigcap_{i=1}^{N} (\phi_i^* \oplus C) \right) = 1$ we first have

$$P_X^n \left[\left| L\left(\overline{\bigcap_{i=1}^{N} \hat{E}_i} \right) - L\left(\overline{\bigcap_{i=1}^{N} (\phi_i^* \oplus C)} \right) \right| > \epsilon \right] \leq P_X^n \left[\left| L\left(\bigcap_{i=1}^{N} \hat{E}_i \right) - L\left(\bigcap_{i=1}^{N} (\phi_i^* \oplus C) \right) \right| > \epsilon \right].$$

This quantity is upper bounded by

$$P_X^n \left[\left| L\left(\bigcap_{i=1}^{N} \hat{E}_i \right) - L\left(\bigcap_{i=1}^{N} (\hat{\phi}_i \oplus C) \right) \right| > \epsilon/2 \right] + P_X^n \left[\left| L\left(\bigcap_{i=1}^{N} (\hat{\phi}_i \oplus C) \right) - L\left(\bigcap_{i=1}^{N} (\phi_i^* \oplus C) \right) \right| > \epsilon/2 \right].$$

For the first term, we have[1]

$$P_X^n\left[\sup_{E_1,E_2,\ldots,E_N}\left|L\left(\bigcap_{i=1}^N E_i\right)-\hat{L}\left(\bigcap_{i=1}^N E_i\right)\right|>\epsilon/4\right]\leq 8s\left(\bigcap_{i=1}^N \mathcal{E}_i,n\right)e^{-\epsilon^2 n/512}.$$

Thus, given $\hat{\phi}_1,\hat{\phi}_2,\ldots,\hat{\phi}_N$, with probability $1-8s\left(\bigcap_{i=1}^N \mathcal{E}_i,n\right)e^{-\epsilon^2 n/128}$, we have

$$L\left(\bigcap_{i=1}^N \hat{E}_i\right)\leq \hat{L}\left(\bigcap_{i=1}^N \hat{E}_i\right)+\epsilon/4\leq \hat{L}\left(\bigcap_{i=1}^N \hat{\phi}_i\oplus C\right)+\epsilon/4$$

where the last step is due to the empirical consistency of $\hat{E}_i$'s. Then simulateneously we have

$$\hat{L}\left(\bigcap_{i=1}^N \hat{\phi}_i\oplus C\right)+\epsilon/4\leq L\left(\bigcap_{i=1}^N \hat{\phi}_i\oplus C\right)+\epsilon/2.$$

Thus we have

$$L\left(\bigcap_{i=1}^N \hat{E}_i\right)\leq L\left(\bigcap_{i=1}^N \hat{\phi}_i\oplus C\right)+\epsilon/2.$$

For the second term, first we have[1]

$$P_X^n\left[\sup_{\phi_1,\phi_2,\ldots,\phi_N}\left|L\left(\bigcap_{i=1}^N(\phi_i\oplus C)\right)-\hat{L}\left(\bigcap_{i=1}^N(\phi_i\oplus C)\right)\right|>\epsilon/4\right]\leq 8s\left(\bigcap_{i=1}^N \mathcal{H}_i,n\right)e^{-\epsilon^2 n/512}.$$

With probability $1-8s\left(\bigcap_{i=1}^N \mathcal{H}_i,n\right)e^{-\epsilon^2 n/512}$, we have

$$L\left(\bigcap_{i=1}^N(\hat{\phi}_i\oplus C)\right)\leq \hat{L}\left(\bigcap_{i=1}^N(\hat{\phi}_i\oplus C)\right)+\epsilon/4\leq \hat{L}\left(\bigcap_{i=1}^N(\phi_i^*\oplus C)\right)+\epsilon/4\leq L\left(\bigcap_{i=1}^N(\phi_i^*\oplus C)\right)+\epsilon/2,$$

where the second inequality is due to the empirical minimization property of ϕ_i^*, i.e., $\hat{L}(\hat{\phi}_i\oplus C)\leq \min_{\phi\in\mathcal{H}_i}\hat{L}(\phi\oplus C)$.

Thus, with probability $1-8\left[s\left(\bigcap_{i=1}^N \mathcal{E}_i,n\right)+s\left(\bigcap_{i=1}^N \mathcal{H}_i,n\right)\right]e^{-\epsilon^2 n/512}$ we simultaneously satisfy the two following conditions

$$L\left(\bigcap_{i=1}^N \hat{E}_i\right)\leq L\left(\bigcap_{i=1}^N \hat{\phi}_i\oplus C\right)+\epsilon/2$$

$$L\left(\bigcap_{i=1}^N(\hat{\phi}_i\oplus C)\right)\leq L\left(\bigcap_{i=1}^N(\phi_i^*\oplus C)\right)+\epsilon/2.$$

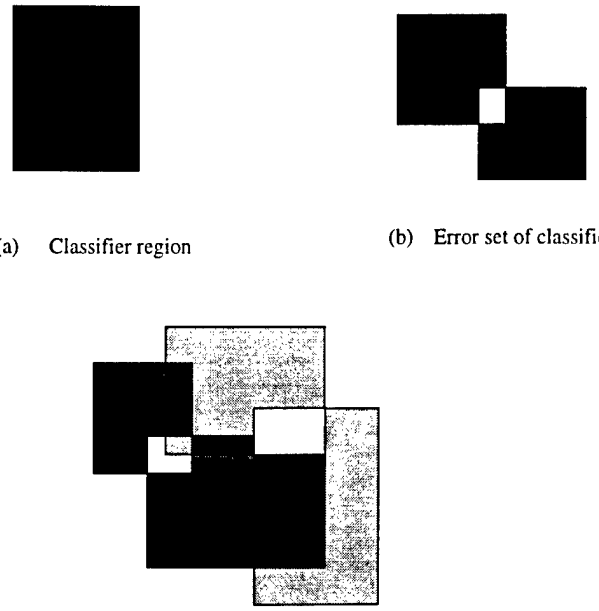
Hence, with the same probability we have

$$\left|L\left(\bigcap_{i=1}^N \hat{E}_i\right)-L\left(\bigcap_{i=1}^N(\hat{\phi}_i\oplus C)\right)\right|<\epsilon,$$

and hence the theorem. □

Corollary 1. Since $L(\tilde{f})\leq L(\tilde{f}_B)\leq \min_{i=1}^N L(\phi_j^*)$, with the same probability as in Theorem 2, we have both the following guarantees

$$P_X^n\left[\left|L\left(\bigcap_{i=1}^N \hat{E}_i\right)-L(\tilde{f}_B)\right|>\epsilon\right]\quad\text{and}\quad P_X^n\left[\left|L\left(\bigcap_{i=1}^N \hat{E}_i\right)-\min_{i=1}^N L(\phi_i^*)\right|>\epsilon\right]$$



(a) Classifier region

(b) Error set of classifier

(c) Intersection of error sets of two classifier

Figure 3. Illustration for Example 3.

under the conditions of the theorem. □

Recall from the previous section that

$$P_X^n\left[L(\hat{\phi}_{\min}) - \min_{i=1}^{N} L(\phi_i^*) > \epsilon\right] \leq \sum_{i=1}^{N} \delta_i = \sum_{i=1}^{N} s(\mathcal{H}_i, n) e^{-\epsilon^2 n/128}.$$

Thus, the guarantee of this method is better under the condition

$$\left[s\left(\bigcap_{i=1}^{N} \mathcal{E}_i, n\right) + s\left(\bigcap_{i=1}^{N} \mathcal{H}_i, n\right)\right] \leq \sum_{i=1}^{N} s(\mathcal{H}_i, n) e^{-3\epsilon^2 n/512}.$$

In general, for two family of sets $\mathcal{A}$ and $\mathcal{B}$, we have

$$s(\mathcal{A} \cap \mathcal{B}, n) \leq s(\mathcal{A}, n) s(\mathcal{B}, n),$$

which make this condition difficult to satisfy in terms of the general upper bound. In practice, $s(\mathcal{A} \cap \mathcal{B}, n)$ could be much smaller.

The applicability of Theorem 2 depends on the choice of $\mathcal{E}_i$'s. We now consider two illustrative examples.

Example 3. Consider that C is a d-rectangle, i.e. $C = \prod_{i=1}^{N} [l_i, h_i]$ for $l_i \leq h_i$, l_i, $h_i \in \Re$ (Fig. 3(a)). Then $\mathcal{H}_i$ can be a set of d-rectangles, and thus $V_{\mathcal{H}_i} \leq 2d$.[1] Since N-fold intersection of d-rectangle is also a d-rectangle, the VC dimension of $\bigcap_{i=1}^{N} \mathcal{H}_i$ is upper bounded by $2d$. Let $\mathcal{R}$ be set of all d-rectangles, and hence $V_{\mathcal{R}} = 2d$. Now we can choose $\mathcal{E}_i = \{C_1 \oplus C_2 : C_1, C_2 \in \mathcal{R}\}$, $i = 1, 2, \ldots, N$, which makes it empirically consistent with any $\hat{\phi}_i \oplus C$. In this case $\bigcap_{i=1}^{N} \mathcal{E}_i$ consists of unions of at most $2N$ rectangles as shown in Fig. 3, and thus its VC dimension is upper bounded by $4Nd$. □

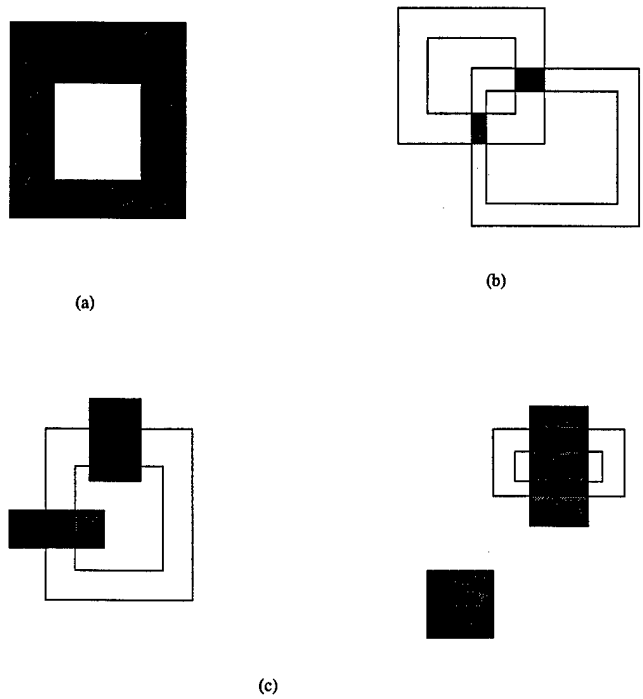


Figure 4. Illustration for Example 4.

Example 4. We now consider a more restrictive and two-dimensional version of Example 3, where we stipulate that $\hat{\phi}_i \subseteq C$ or $C \subseteq \hat{\phi}_i$ for $i = 1, 2, \ldots, N$. Such classifiers can be easily realized by computing the largest or smallest rectangles that include all positive examples. In this case the error sets are "rectangular rings" as shown in Fig. 4(a). As result, the intersection of error sets consists of at most two rectangles as shown in Fig. 4(b) and (c), assuming the $\hat{\phi}_i$'s are all distinct. Thus for this case, we have $V_{\mathcal{H}_i} \leq 4$. The VC dimension of the error sets is no more than 8, and that of their intersection is also no more than 8. In this case the condition of the Theorem 2 is easily satisfied. □

5. CONCLUSIONS

We presented two methods for fusing classifiers so that the fused system provides better performance guarantees than the best classifier. Under additional conditions of deterministic classes and consistent error set estimates, we showed that the fused system provides better guarantees than any subset of the classifiers. There are several avenues for future research. First, extensions of the second method to more general cases such as probabilistic classes and regression estimation, will be of interest. Second, the notion of metafusers[17] that combine the fusers is very appealing. For the first method based on isolation property, metafusers do not offer much more than what is feasible by fusing all classifiers, i.e. the fused system simply retains the performance of the best classifier. On the other hand, the second method might provide a performance significantly better than the best classifier, and hence a metafuser might reduce error below the levels possible by individual fusers. Such reduction is possible only when accurate estimation of error sets can be carried out efficiently. It would of future interest to investigate the performance trade-offs involved in such metafuser design.

ACKNOWLEDGMENTS

This research is sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp., the Seed Money Program of Oak Ridge National Laboratory, and the Office of Naval Research under order N00014-96-F-0415.

REFERENCES

1. L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.
2. P. K. Varshney, *Distributed Detection and Data Fusion*, Springer-Verlag, 1996.
3. C. K. Chow, "Statistical independence and threshold functions," *IEEE Trans. Electronic Computers* **EC-16**, pp. 66–68, 1965.
4. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
5. V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.
6. L. G. Valiant, "A theory of the learnable," *Communications of the ACM* **27**(11), pp. 1134–1142, 1984.
7. M. Vidyasagar, *A theory of Learning and Generalization*, Springer-Verlag, New York, 1997.
8. B. Grofman and G. Owen, eds., *Information Pooling and Group Decision Making*, Jai Press Inc., Greenwich, Connecticut, 1986.
9. N. S. V. Rao and S. S. Iyengar, "Distributed decision fusion under unknown distributions," *Optical Engineering* **35**(3), pp. 617–624, 1996.
10. N. S. V. Rao, "Distributed decision fusion using empirical estimation," *IEEE Transactions on Aerospace and Electronic Systems* **33**(4), pp. 1106–1114, 1996.
11. N. S. V. Rao, "Fusion methods for multiple sensor systems with unknown error densities," *Journal of Franklin Institute* **331B**(5), pp. 509–530, 1994.
12. N. S. V. Rao, "Nadaraya-Watson estimator for sensor fusion," *Optical Engineering* **36**(3), pp. 642–647, 1997.
13. N. S. V. Rao, "Vector space methods for sensor fusion problems," *Optical Engineering* **37**(2), 1998. in press.
14. N. S. V. Rao and E. M. Oblow, "Majority and location-based fusers for PAC concept learners," *IEEE Trans. on Syst., Man and Cybernetics* **24**(5), pp. 713–727, 1994.
15. N. S. V. Rao and E. M. Oblow, "N-learners problem: System of PAC learners," in *Computational Learning Theory and Natural Learning Systems, Vol IV: Making Learning Practical*, pp. 189–210, MIT Press, 1997.
16. N. S. V. Rao, E. M. Oblow, C. W. Glover, and G. E. Liepins, "N-learners problem: Fusion of concepts," *IEEE Transactions on Systems, Man and Cybernetics* **24**(2), pp. 319–327, 1994.
17. N. S. V. Rao, "Function estimation using multiple PAC estimators," *Machine Learning* , 1998. submitted.

Report Number (14) ORNL/CP--97084
CONF-980412--

Publ. Date (11) 199804
Sponsor Code (18) DOE/ER; DOD, XF
JC Category (19) UC-400; UC-000, DOE/ER

DOE